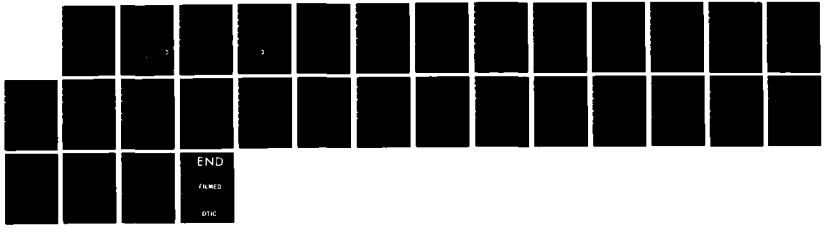
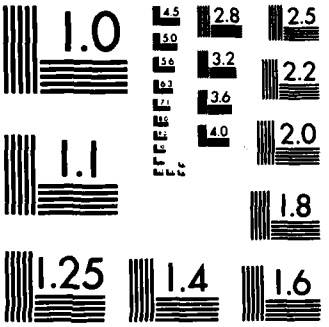
AD-A155 006 AUTOMATIC SPEAKER RECOGNITION FOR MILITARY APPLICATIONS: APPLICATIONS SURVEY AND OPERATIONAL REQUIREMENTS(U) NAVAL RESEARCH LAB WASHINGTON DC S S EVERETT 07 MAY 85 NRL-MR-5545 1/1

UNCLASSIFIED F/G 17/2 NL





NRL Memorandum Report 5545

Automatic Speaker Recognition For Military Applications:

Applications Survey and Operational Requirements

S. S. EVERETT

Communication System Engineering Branch
Information Technology Division

May 7, 1985

NAVAL RESEARCH LABORATORY
Washington, D.C.

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b RESTRICTIVE MARKINGS
2a SECURITY CLASSIFICATION AUTHORITY	3 DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b DECLASSIFICATION / DOWNGRADING SCHEDULE	
4 PERFORMING ORGANIZATION REPORT NUMBER(S) NRL Memorandum Report 5545	5 MONITORING ORGANIZATION REPORT NUMBER(S)
6a NAME OF PERFORMING ORGANIZATION Naval Research Laboratory 6b OFFICE SYMBOL (*If applicable*) Code 7526	7a NAME OF MONITORING ORGANIZATION
6c ADDRESS (*City, State, and ZIP Code*) Washington, DC 20375-5000	7b ADDRESS (*City, State, and ZIP Code*)
8a NAME OF FUNDING / SPONSORING ORGANIZATION Office of Naval Research 8b OFFICE SYMBOL (*If applicable*)	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c ADDRESS (*City, State, and ZIP Code*) Arlington, VA 22217	10 SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO: 61153N; PROJECT NO:; TASK NO: RR021 05-42; WORK UNIT ACCESSION NO: DN480-556

11 TITLE (*Include Security Classification*) Automatic Speaker Recognition for Military Applications: Applications Survey and Operational Requirements

12 PERSONAL AUTHOR(S)
Everett, S.S.

13a TYPE OF REPORT	13b TIME COVERED	14 DATE OF REPORT (*Year, Month, Day*)	15 PAGE COUNT
Interim	FROM _____ TO _____	1985 May 7	30

16 SUPPLEMENTARY NOTATION

17 COSATI CODES			18 SUBJECT TERMS (*Continue on reverse if necessary and identify by block number*)
FIELD	GROUP	SUB-GROUP	Speech processing; Communication security; Voice processing; Speaker recognition; Access control

19 ABSTRACT (*Continue on reverse if necessary and identify by block number*)

Automatic speaker recognition (ASR) systems, capable of identifying a person based on voice input alone, have a wide range of potential applications in Navy environments. They could provide access control to restricted areas, equipment and information. They could verify the identity of users of various communication channels, or verify computer users through terminals accepting voice input. This report presents the results of a survey conducted to determine which type of application offers the best potential use of ASR for the Navy. The survey results also give a rough indication of the operational requirements of ASR systems in Naval environments. In designing an ASR system for a particular application there are a number of factors regarding the user and the environment that must be taken into consideration. These factors, and their impact on the operational requirements, are discussed in detail. The appendix of this report outlines the various approaches to ASR and the major ASR systems described in the literature.

20 DISTRIBUTION / AVAILABILITY OF ABSTRACT	21 ABSTRACT SECURITY CLASSIFICATION
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	UNCLASSIFIED
22a NAME OF RESPONSIBLE INDIVIDUAL S. S. Everett	22b TELEPHONE (*Include Area Code*) (202) 767-2116 22c OFFICE SYMBOL Code 7526

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted
All other editions are obsolete

SECURITY CLASSIFICATION OF THIS PAGE

CONTENTS

AUTOMATIC SPEAKER RECOGNITION FOR MILITARY APPLICATIONS:
Applications Survey and Operational Requirements

I. BACKGROUND

In the last two decades considerable research has been performed aimed at developing methods of recognizing speakers automatically based on voice input alone. Such systems could be of considerable benefit in numerous situations in both civilian and military environments. Potential Navy applications include access control for restricted areas or information, communication security and verification of computer users through terminals accepting voice input.

There are two primary functions of automatic speaker recognition. The first is speaker verification, where the system either accepts or rejects the identity claim made by the user based on his or her voice characteristics. The second is speaker identification, where the system determines which speaker from a known set best matches the unknown input voice. In this paper the use of the term automatic speaker recognition (ASR) indicates both speaker verification and speaker identification.

Numerous ASR systems have been developed, but all follow the same basic steps (see Fig. 1). First, reference parameters, or templates, are created for each speaker based on one or more training utterances. Parameters derived from unknown test utterances are then compared with these templates. The degree of similarity between the two sets of parameters is evaluated using some form of distance measure. In speaker verification systems the user's identity claim calls forth the reference data for that speaker. The claim is accepted if the distance between the test and training parameter sets is below some pre-determined threshold. In speaker identification systems no identity

Manuscript approved December 13, 1984.

claim is made by the user. Instead the input utterance is compared with the stored templates for all the speakers in the set, and the speaker is identified as the one whose reference data most closely matches the input speech. Some identification systems allow a "none of the above" choice if the distances for all speakers exceed a given threshold.

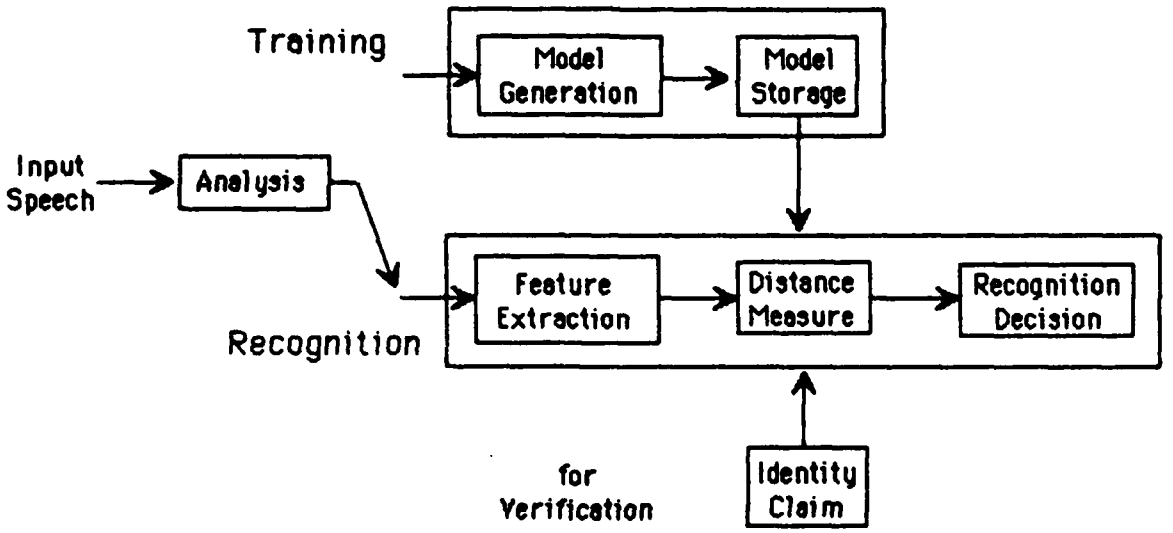


Figure 1. Block diagram of basic ASR system methodology.

A majority of the research on ASR has been concentrated on the question of which parameters yield the best results, and what is the best method of deriving them. The numerous approaches to ASR can be divided into two basic categories. The first is based on parameters derived through linear predictive analysis of the speech waveform (LPC), such as reflection coefficients, autocorrelation coefficients, log area ratios, etc. [Atal, 1974; Sambur, 1976(a, b); Markel, 1977; and others]. The second approach, sometimes called "statistical", is based on parameters measured directly from the speech waveform, such as pitch, intensity, spectral characteristics, etc. [Luck, 1966; Furui et al, 1972; Rosenberg, 1976(b); and others]. Neither approach has been shown to be clearly superior to the other over all. There is also a

sort of hybrid approach to ASR that involves the use of both LPC-based and statistical parameters [Furui, 1981; Mohankrishnan et al, 1982]. For a description and discussion of specific systems or approaches see the Appendix of this report.

II. APPLICATIONS SURVEY

As mentioned above there is a wide variety of potential applications for ASR in Navy environments. In addition, there are a number of factors to consider in specifying an ASR system for a given application. In order to determine which applications of ASR would provide the greatest benefit and what the operational requirements in those applications would be, a written survey was conducted. Roughly 50 questionnaires were distributed, with approximately 40% returned.

Reaction to the survey was positive, and indicated that if the accuracy of ASR systems could meet the stringent requirements of the military, their use could be very widespread. The primary findings of the survey are summarized in Table 1.

The potential applications presented in the questionnaire can be divided into three categories: (1) access control to restricted areas and/or equipment, (2) communication security, including both the verification of channel users and the monitoring of channel activity, and (3) computer security. As can be seen from the table, there is interest in ASR for all of these areas, but it is strongest for access control and communication security.

Though the idea of having one ASR system to serve in all applications is attractive, it is not practical -- different applications will require different ASR capabilities. There are so many variables in terms of the environment, the role of the system, and the user, that it would be impossible

to design an efficient ASR system that could be used satisfactorily in every situation. Before a system is designed for a specific application many factors need to be considered. The survey was designed to assess some of these factors for the applications presented. The following pages contain detailed discussions of these considerations and the survey results associated with some of them.

Table 1. Summary of ASR applications survey results.

Question or Category	Number of responses
Potential Areas of Application	
Access control	
Areas	XXXXXXXXXX
Equipment	XXXXXXXX
Communication security	
Verify users	XXXXXXXXX
Monitor channel	XXXXXXX
Computer security	XXXXXXX
Typical Noise Level of Application Environment	
Low - may speak normally	XXXXXXXXX
Moderate - must raise voice slightly	XXXXXXXXXX
High - must speak loudly	XXXXXX
Extreme - must shout to be heard	X
Number of Speakers on File	
Less than 5	X
5 to 10	XXX
10 to 20	XXXX
20 to 50	XXXXXX
Over 50	XXXXXXXXX
Required Speed of Recognition Process	
Speed not critical	XXXXX
5 seconds	XXXXXX
2 seconds	XXXXX
1 second	XXXX
.5 second	X
Maximum Tolerable Error Rate	
Type I Error (True speaker rejection)	
1 in 10	X
1 in 20	XXXXX
1 in 50	
1 in 100	XXXXXX
1 in 1000	XXXXXX
Type II Error (Impostor acceptance)	
1 in 10	XX
1 in 100	XXXXXX
1 in 1000	XXXXXX
1 in 1,000,000	XXXX
Recognition Method	
Text-independent recognition	XXXXXXXX
Text-dependent recognition	XXXXXXXXXXXXX

III. ASR SYSTEM DESIGN CONSIDERATIONS

All of the issues discussed below, and undoubtedly others which are not mentioned, should be addressed in determining what would be required of an ASR system in a given application.

A. Speaker Identification or Speaker Verification This and the question of text-dependent or text-independent operation (see B below) are probably the two most basic issues in designing an ASR system for a particular application. The question of identification versus verification is related primarily to the function that the system is to perform. If, for example, the system were to be used for access control, speaker verification would be most practical since all that is needed is a Yes/No response to the user's identity claim. If, on the other hand, the system were to be used in the monitoring of communication channel activity, speaker identification would be the obvious choice for determining which of a known set of speakers is using the line.

In theory speaker verification is the easier and the more accurate because it involves a single binary decision, whereas speaker identification involves comparing the unknown input speaker to the stored templates for all speakers in the set and choosing the best match.

B. Text-Dependent or Text-Independent Operation ASR systems can be either text-dependent, where the parameters are obtained from particular words or phrases, or text-independent, where the text is unconstrained and the parameters are obtained by averaging over some time interval. Most approaches can be implemented in either kind of system. Since the accuracy scores for both text-dependent and text-independent systems are comparable, this choice is also related primarily to the function of the ASR system in the given application.

Text-independent systems offer obvious benefits in situations where the speaker is not cooperative or is not aware of the recognition process. In situations such as access control where the speaker is assumed to be cooperative, text-dependency offers a measure of additional security in that the speaker must provide the specific word or phrase to which the system has been trained. This also means that the system is less vulnerable to compromise by tape-recorded material, since one would need a recording of the required word or phrase.

Many applications may require the ability to recognize speakers based on only short segments of input speech -- it is not always feasible to wait 20 or 30 seconds before making a decision. As a rule, text-dependent systems require less input speech to reach a decision than do text-independent systems. However, text-dependent systems also require fairly accurate time alignment to allow exact comparison of the input and the reference parameters. This is often accomplished through non-linear time deformation (warping) using pattern-matching techniques. Warping is not necessary in a text-independent system because the parameters are simply averaged over some period of time, and then compared with the stored templates.

The survey results show a definite preference for text-dependent speaker recognition, particularly in access control and computer security applications. For communication security applications the preference is less clear, except in monitoring activities, which naturally require text-independent recognition. Of the two approaches, text-dependent recognition is probably the easier because the phonetic content of the input is known. However, the tricky problem of accurate end-point detection is much more important with this approach than with text-independent recognition.

C. Recognition at Transmitter vs. Recognition at Receiver The question of whether to perform speaker recognition at the transmitter or at the receiver is fundamental to the design of a system that is to operate over communication channels. The greatest benefits of performing recognition at the transmitter are that in this way the ASR system does not have to cope with the transmission characteristics of the channel itself, and the entire speech waveform is available, thus allowing the use of whatever analysis or measurement technique is desired. Recognition at the receiver, on the other hand, must be performed after the signal has probably been corrupted by the transmission channel, and is restricted to using only the data that is sent.

There are two major drawbacks to performing the recognition at the transmitter if the equipment is to be used in the field. First, the transmitter in such a situation is much more vulnerable than is the receiver, and therefore the system is more subject to compromise. Second, in a tactical or emergency situation it may not be wise to deny the user access to the channel if he or she cannot be recognized. Recognition at the receiver would be less subject to compromise, but would also allow anyone access to the channel. This access could be critical in times of emergency, and would still allow the receiver to monitor possible impostors.

In the survey responses there was a unanimous recommendation that speaker recognition be performed at the receiver rather than at the transmitter unless the goal is to restrict access to the communication channel itself. If a transmitter were compromised this approach would allow the receiver to be aware of the situation without revealing this knowledge to the transmitter. However, if the transmitter site is known to be secure, and if there are sufficient alternative procedures in case of emergency, higher accuracy could probably be obtained by performing the speaker recognition at the transmitter

where the full voice signal is available and the transmission characteristics of the channel itself are not a factor.

D. Input Speech Quality There are many situations where an ASR system would be required to operate using degraded or distorted input speech. These include telephone speech, vocoded speech, military communication channels, and situations where the user is breathing a special air mixture. The quality of the input speech will also be affected if the user is wearing headgear, such as an oxygen mask, or is subject to high G forces. ASR systems designed to operate with distorted or vocoded input speech may require different parameters than those designed to work under more ideal conditions.

Several ASR systems described in the literature are capable of performing recognition with slightly degraded speech. Some of these were developed using actual telephone speech; other researchers have used "telephone quality" speech -- usually band-limited and contaminated with some sort of artificial noise. It has been shown, however, that results obtained using speech contaminated with white noise cannot be extrapolated to indicate performance over actual telephone channels [Shridhar and Baraniecki, 1979]. Further discussion of research on ASR using telephone speech may be found in the Appendix.

Almost no research has been done on ASR using speech from other communication channels such as narrowband channels where the speech is processed or encoded. The analyses performed to allow bandwidth compression of the speech signal frequently remove or distort certain characteristics of the original speech. In the development of these techniques care has been taken to preserve the quality and intelligibility of the processed speech. However, it is not known what effects, if any, the processing has on the portions of the speech signal relevant to speaker identity.

VI. REFERENCES

B. S. Atal, 1972
"Automatic Speaker Recognition Based on Pitch Contours," Jour. Acoust. Soc. Am., vol. 52, pp. 1687-1697.

B. S. Atal, 1974
"Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," Jour. Acoust. Soc. Am., vol. 55, pp. 1304-1312.

B. S. Atal, 1976
"Automatic Recognition of Speakers from Their Voices", Proceedings IEEE, vol. 64, pp. 460-475.

B. S. Atal and S. L. Hanauer, 1971
"Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," Jour. coust. Soc. Am., vol. 50, pp 637-655.

C. J. Atkinson, 1952
"A Study of Vocal Responses During Controlled Aural Stimulation," Jour. Speech and Hearing Disorders, vol. 17, pp. 419-426.

M. Baraniecki and M. Shridhar, 1980
"A Speaker Verification Algorithm for Speech Utterances Corrupted by Noise with Unknown Statistics," Proceedings IEEE 1980 Intl. Conf. Acoustics, Speech and Signal Proc., pp. 904-907.

H. M. Dante and V. V. S. Sarma, 1979
"Automatic Speaker Identification for a Large Population," IEEE Trans. Acoustics, Speech and Signal Proc., vol. ASSP-27, pp. 255-263.

G. R. Doddington, R. E. Helms, and B. M. Hydrick, 1976
"Speaker Verification III," RADC-TR-76-262, Final Technical Rept. for Rome Air Development Ctr., Aug. 1976. AD B014720L

S. Furui, 1981
"Comparison of Speaker Recognition Methods Using Statistical Features and Dynamic Features," IEEE Trans. Acoustics, Speech and Signal Proc., vol. ASSP-29, pp. 342-350.

S. Furui, F. Itakura and S. Saito, 1972
"Talker Recognition by Longtime Averaged Speech Spectrum," Electronics and Communications in Japan, vol 55-A, No. 10, pp. 54-61.

S. Furui and A. E. Rosenberg, 1980
"Experimental Studies in a New Automatic Speaker Verification System Using Telephone Speech," Proceedings IEEE 1980 Intl. Conf. Acoustics, Speech and Signal Proc., pp. 1060-1062.

G. R. Griffin and C. E. Williams, 1984
"The Effect of Workload on the Vocal Utterances of Individuals Performing a Simulated Flight Task," Proceedings Am. Voice I/O Soc., Sept. 1984, paper 20.

Secrest and Helms [1977] used telephone speech in their development of a text-dependent speaker verification system which was based on spectral error parameters. A three-prong compensation technique was implemented to compensate for the effects of the telephone channel. First, the data was normalized by measuring the energy in the input signal during periods of silence. Second, a band-limited spectrum (400 to 2620 Hz) was used for time registration of the input phrases to minimize problems caused by the amplitude distortion present in the channel. Third, pitch period information was included to provide additional speaker discrimination, since pitch is relatively insensitive to the bandpass characteristics of the telephone channel. A limited test was conducted with 16 speakers using random phrases of 4 mono-syllabic words (from a set of 32 allowable phrases). Error rates reported were 1% true speaker rejection and 1% false speaker acceptance.

Mohankrishnan et al [1982] also showed significant improvements in speaker identification accuracy gained by combining log area ratios with the output of a bank of 16 band-pass filters. Individually, each approach yielded an accuracy of 93.2%. This was raised to 97.7% when the two were combined. Another combination using the inverse filter spectrum plus the band-pass filters raised scores from 97.4% and 93.2% respectively to 98.3%.

By combining parameters in this way, each parameter can serve as a sort of double-check, or confirmation, of the other. Also, in cases where the results for one parameter are ambiguous, the second can provide clarification. This approach should be particularly effective if the parameters used are not highly correlated, and thus measure different aspects of the speech signal.

D. ASR Using Telephone Speech Sambur [1976] used actual telephone speech in evaluating the effectiveness of orthogonal linear predictive parameters for speaker identification. Initial results showed an accuracy rate of about 85%, as opposed to more than 90% for the same system when using high-quality input speech. However, it was shown that the channel distortion introduced by the telephone channel affected primarily the four least significant reflection coefficients. For this reason a 14^{th} order analysis was performed, from which only the first 10 coefficients were used in the identification process. This raised accuracy scores to roughly 95%.

Baraniecki and Shridhar [1980] also investigated the use of orthogonal linear predictive parameters with noisy telephone speech. Their initial verification accuracy was around 45% using a 12^{th} order LPC analysis. By implementing a noise cancellation technique consisting of a self-tuning non-recursive filter, the accuracy scores were raised to more than 96%.

It should be noted that, although both pitch and intensity have been used successfully in ASR systems, these parameters are highly susceptible to stress on the speaker and to efforts to mimic or to disguise the voice. For this reason, they are probably best used in combination with other parameters.

Cepstral measurements have also been used by several investigators. Luck [1966] described a speaker verification system that used cepstral measurements taken from short vowel segments in a fixed utterance. Accuracy scores were given as ranging from 87% to 94%, even under conditions of intentional (but untrained) mimicry. Furui et al [1972] used cepstral measurements derived from the long-term average of the power spectrum. Accuracies for text-dependent speaker identification and verification were 91% and 94% respectively.

Other non-LPC based parameters that have shown potential for use in ASR systems include the glottal source spectrum slope [Wolf, 1972], formant frequencies [Sambur, 1973], the long- and short-term speech spectra [Höfker and Jesorsky, 1979; Mohankrishnan et al, 1982], and the voice onset time (VOT) of stop consonants [Wolf, 1972; Sambur, 1973].

C. Hybrid Approaches Although reasonable accuracy rates can be obtained using either LPC or statistical parameters, it has been shown that even higher accuracy can be obtained by using more than one parameter in a given system. Research on systems of this type has been scarce, but the results are highly promising. Furui [1981] combined "dynamic" parameters (time functions of the log area ratios and the fundamental frequency) with "statistical" parameters (extracted from the longtime averaged spectrum). This combination produced a 50% drop in error rate as compared with the results obtained from either parameter alone. Identification accuracy was raised from approximately 97.6% (dynamic) and 98.1% (statistical) to 99.1% with the combined parameters.

B. Non-LPC Based Approaches As early as 1963 efforts were being made to develop ASR systems. Because computational capabilities were limited, investigators turned to quantized spectrographic information [Pruzansky and Matthews, 1964] and analog filter banks [Hargreaves and Starkweather, 1963]. Though slow and tedious, these methods yielded speaker identification accuracies around 90% in both investigations. Doddington et al [1976] also based their speaker verification system on the output of an analog filter bank; accuracy was reported as less than 1% speaker rejection and less than 2% imposter acceptance.

As computational capabilities have expanded, more complex and sophisticated methods of ASR have been developed. A variety of non-LPC based parameters have been investigated, with varying degrees of success. The two most widely-used parameters are the intensity of the voice and the speaker's pitch, or fundamental frequency. Atal [1972] investigated text-dependent speaker identification based on the pitch contours of sentence-long utterances; accuracy was reported as 97% using high-quality input speech. Rosenberg [1976(b)] used similar pitch contours and combined them with intensity contours, yielding an overall accuracy of 91% for text-dependent speaker verification using telephone speech. Wolf [1972] found the fundamental frequencies of certain (manually located) vowel segments to be "useful parameters." Markel et al [1977] found that the fundamental frequency yielded a relatively high variance ratio (the ratio of the intra-speaker variance to the inter-speaker variance), indicating a good potential for use in ASR systems; the standard deviation of the gain (intensity) deviation showed considerably less promise. No accuracy scores were given.

coefficients. The averaging of up to 1000 voiced frames (about 70 seconds of speech) yielded accuracy scores up to approximately 98% [Markel et al, 1977, Markel, 1978].

Furui and Rosenberg [1980] developed a system using cepstral coefficients derived from the linear predictive coefficients. Analysis was performed on fixed sentence-long utterances, resulting in a feature contour as a function of time. This text-dependent speaker verification system operated on both high-quality and telephone speech, and yielded average accuracy rates over 99% in all conditions.

Numerous other investigators have applied the principles of LPC to speaker recognition using a variety of parameters and distance measures [Pfeifer, 1977; Dante and Sarma, 1979; Ney, 1981; Schwartz et al, 1982]. Reported identification accuracies average slightly more than 90%.

In 1976 Sambur pioneered the use of orthogonal linear predictive parameters for speaker recognition. These orthogonal parameters, which are obtained from a linear transformation of the linear prediction parameters, are attractive for two reasons. First, it can be shown that the orthogonal parameters are essentially independent of the linguistic content of the utterance but are highly indicative of speaker characteristics. Second, they do not require any time normalization procedures because they are averaged across an entire utterance, thus making them particularly well suited for use in text-independent systems. Accuracy scores over 99% were reported for text-dependent speaker identification and verification using a 12^{th} order LPC analysis. Text-independent verification scores were around 94% [Sambur, 1976(a), 1976(b)].

V. APPENDIX

The following paragraphs contain brief discussions of some of the important work done on the various approaches to automatic speaker recognition (ASR). It should be noted that the scores quoted in the following discussions are those published by the respective investigators. Since there is no standard method of evaluating speaker recognition systems these scores do not permit direct comparison of the different systems and approaches. They should be interpreted only as rough indications of the level of performance of the systems and not as absolute ratings.

A. LPC-Based Approaches Linear predictive coding (LPC) has become very widely used in speech analysis and synthesis since its introduction in 1971 [Atal & Hanauer, 1971]. These principles have also been applied to ASR by several different researchers. The parameters investigated have included reflection coefficients, log-area ratios, autocorrelation coefficients, clipped autocorrelation coefficients, and other parameters based on the linear predictive analysis of the speech signal.

In 1974 Atal investigated text-dependent ASR using linear predictive coefficients and other parameters derived from them, such as the impulse response function, the autocorrelation function, the area function and the cepstrum function. With only 50 msec of input speech a 12^{th} order LPC analysis yielded average identification accuracies ranging from 57.0% for the area function to 70.3% for the cepstrum coefficient. Increasing the sample duration to 0.5 sec reportedly raised the accuracy of the cepstrum function to 98% [Atal, 1974].

Markel obtained similar accuracy for text-independent speaker identification by using the long-term averages of the reflection

monitoring of channel activity. With the future development of computer terminals capable of accepting voice input, ASR will become an ideal way of verifying computer users and controlling access to stored information.

There are many environmental and user variables that must be considered in determining the operational requirements of an ASR system for a given application. The survey results presented in this report indicate that these requirements for Naval applications include maximum error rates of .1 to 1%, the ability to store templates for 50 or more speakers at a time, and the ability to perform ASR in moderate noise environments or with processed or vocoded input speech.

Though there are no ASR systems currently available that can meet all of these requirements they are not unreasonable performance goals for ASR systems to be developed in the near future. The increasing availability of fast inexpensive computational capabilities, including single-chip speech analysis systems, and our better understanding of the human voice are making reliable real-time ASR systems more and more feasible. It would seem that now is the ideal time to actively pursue a comprehensive program aimed at developing automatic speaker recognition technology for military use. With careful evaluation of application environments and their operational requirements, the Navy could realize significant benefit and savings from the implementation of this technology.

If the ASR system has not been trained using a stressed voice, it may have difficulty recognizing the speaker under these conditions. Unfortunately, it is often hard to simulate such stresses for training purposes.

L. What Happens if the User is Rejected? To operate successfully, an ASR system must be "friendly" to the user, particularly in cases when the speaker is rejected or the system is unable to make an identification. Would the user know if he or she had been rejected? Would he be prompted to try again? If so, how many tries would be allowed? Would there be the possibility of overriding the recognizer? In a system that operates over a communication channel would the user be denied access to the channel, even in a time of emergency, if verification or identification could not be obtained? Or would access to the channel be open, with possible impostors or unauthorized users flagged only at the receiver? These questions are fundamental to the design of a friendly and efficient system.

Survey responses on this subject were mixed. The exact method of handling rejections is very application dependent, but most participants felt the system should reset in 10-20 seconds or prompt the user to try again, allowing up to three tries. However, recommendations ranged from denying access and alerting security personnel without giving the user a second chance, to allowing the user to pass anyway and merely keeping a record of the rejection.

IV. CONCLUSION

ASR offers the Navy great potential for increased security and personnel verification capabilities in numerous situations. Through the use of automatic speaker verification, human guards would no longer be needed for controlling access to restricted areas or equipment. ASR would also be a valuable tool for verifying users of communication channels and in the

training procedures can be particularly troublesome with text-dependent systems because they need to be completely retrained each time the word or phrase is changed. Such limitations should be considered in determining the requirements of a given application.

J. The User One aspect of ASR systems and applications that is often overlooked is the human side of the problem, namely the user. The effective implementation of any ASR system, especially a speaker verification system, requires the complete confidence and cooperation of the people who will be using it. If the system is not user-friendly, this cooperation will be difficult to maintain, and the performance of the system will consequently drop. Attention should be paid to the type of person who will be using the system, and to their backgrounds, attitudes, and abilities. Whether or not the person will be required to perform other simultaneous tasks should also be considered. Any system, therefore, must be designed with both the application and the user in mind. This is particularly important in cases where the user is apt to be under stress.

K. Stresses on the User Peoples' voices change markedly under conditions of physical and/or emotional stress. Though little research has been done on the effects of these voice changes on the performance of ASR systems, other investigations have indicated that stress produces changes in the properties of speech sounds [Mosko et al, 1983; Griffin and Williams, 1984]. These changes include increased pitch and vowel formant frequencies as well as alterations in the waveform of the glottal pulse and in the amplitude of turbulence noise for stop consonants. In addition, speakers tend to be less consistent in their pronounciations of words in stressful situations. These differences are not predictable, however, because some speakers are more affected by stress than others.

Changes in peoples' voices over a long period of time will eventually affect the performance of an ASR system. For example, in one study recognition accuracy fell from 96% to 52% when the interval between the test and training utterances was increased from 3 days to 3 months [Furui et al, 1972]. For this reason it is necessary to update the reference parameters periodically. Text-dependent systems are effectively updated whenever the system is retrained for a new word or phrase. Updating of text-independent systems, though, has to be handled differently. The time between updates can vary from a number of weeks to a number of years depending on the parameters used by the system and on the individual speaker.

Various methods of automatically updating the templates have been suggested. One is to incorporate every N^{th} test utterance into the templates for that speaker, thereby keeping sort of a running profile of the speaker [Rosenberg, 1976(a)]. Another method, for use in a speaker verification system, is to update the templates with each rejected utterance so the speaker will not be rejected for the same problem over and over again [Höfker and Jesorsky, 1979]. Both of these methods are acceptable for experimental purposes, but have definite drawbacks in practical applications. Automatic updating of reference parameters is especially risky when the ASR system is to be used for security purposes -- one would not want to be updating the templates with an impostor's voice. If the users are at remote sites, the identity of the speaker should be positively confirmed before updating the reference templates.

I. Amount of Training Required Different approaches to ASR require different amounts of training to develop the reference templates, but all systems perform better if the reference templates are formed over a period of days or weeks, rather than in one sitting. However, it is not always feasible to spread the training out over a long period of time. Long drawn-out

This would require more hardware, and would make the verification procedure more complex, but would raise the level of security of the system.

G. Number of Speakers Survey results indicate that ASR systems capable of handling at least 20 speakers would be most useful in military applications. For speaker verification systems, the number of speakers the system can support is limited only by the memory available. Speaker identification systems, on the other hand, become unwieldly with a large number of speakers since the input speech has to be compared to the reference templates for every speaker. Accuracy of a speaker identification system also drops when the speaker set is large. If an application requires speaker identification with a very large set of speakers, some sort of sequential decision strategy may be needed to help speed the identification process and increase the accuracy of the system.

H. Voice Variability in Individual Speakers Perhaps the major barrier to high-accuracy ASR systems is the fact that peoples' voices vary over time. Two utterances of the same word by the same speaker at two different times will never be identical; as the time between the utterances is increased they will differ even more. The health of the speaker (being over-tired, or having a cold, a sore throat, etc.) can also cause marked variations in the speech. However, by creating composite reference templates from several different training sessions separated by a period of days or weeks, the ASR system can be made much less sensitive to this type of variation. Obviously, the more representative the templates are of the day-to-day variations in a particular speaker's voice, the better the system will perform. There is, however, a practical limit to the length of time the training can span -- imagine having to wait six weeks for access to your new office because it is in a restricted area!

different ways. For example, weaker fricatives (/f/, /th/) are easily overwhelmed by wideband noise, whereas vowel sounds remain intelligible even when the background noise is louder than the speech. For this reason code words for text-dependent ASR systems operating in noisy environments must be chosen very carefully.

Survey responses regarding noise level indicate that although there are a number of military platforms with very high noise levels, a majority of the potential Navy applications for ASR have low to moderate noise levels.

F. Accuracy Requirements The performance accuracy required of an ASR system will depend on the specific application. Access control, for example, would probably require higher accuracy than would the monitoring of communication channels. Speaker verification systems, such as would be used for access control, can be biased in favor of false rejection, thereby lowering the possibility of accepting an impostor at the expense of rejecting an occasional true speaker.

Little data is available on the vulnerability of ASR systems to trained professional mimics or close family members. It has been shown that, in general, those parameters which reflect actual physiological characteristics of the speaker (vocal tract measurements, glottal characteristics, etc.) are less susceptible to mimicry than those parameters which reflect learned or behavioral characteristics (pitch, intensity, timing, etc.). However, physiological characteristics of related persons could be very similar, particularly in the case of identical twins.

The survey participants definitely feel accuracy is more important than speed for the applications mentioned. For situations requiring very high accuracy, ASR could easily be augmented with some other form of personnel identification (fingerprints, magnetically coded badges, ID numbers, etc.).

Special air mixtures, such as those used by divers, in submarines or high-altitude jets, etc., can have marked effects on speech quality because the gaseous makeup of the air affects the resonance characteristics of the vocal tract and consequently the spectrum of the speech. Air mixtures containing relatively high proportions of helium, for example, make everyone sound like Donald Duck due to an upward shift of the formant frequencies. For this reason it is imperative that templates be generated using the same air mixture as in the actual application environment. In addition, special identification parameters and parameter extraction methods are probably required for situations of this type, though little research has been done in this area.

E. Operating Environment Another area of concern is that of noisy environments where the input speech is contaminated by background noise. If the noise is relatively consistent, and is included in the generation of the reference templates, then low to moderate noise levels may still allow adequate ASR accuracy. With high noise levels, however, it is necessary to provide some sort of noise reduction (special microphones, spectral subtraction preprocessing [Kang and Everett, 1982], etc.) prior to performing the speaker recognition. This can also affect the quality of the input speech. Perhaps a solution to this problem lies in better methods of reducing noise interference rather than in trying to perform ASR on noise-contaminated speech.

In all cases it is important that the training or template generation be done under conditions as close as possible to the actual environment because people speak differently in noise than in a quiet environment [Atkinson, 1952; Siegel and Pick, 1974]. Specifically, noisy environments cause people to raise the pitch of their voices and to speaker louder in an effort to overcome the background noise. In addition, noise affects different speech sounds in

W. A. Hargreaves and J. A. Starkweather, 1963
"Recognition of Speaker Identity", Language and Speech, vol. 8, pp. 63-67.

A. Higgins and J. Naylor, 1984
Final Report on Contract N00014-84-C-2130, ITT Defense Comm. Div., San Diego, CA, July 1984.

U. Höfker and P. Jesorsky, 1979
"Structure and Performance of an On-Line Speaker Verification System," Proceedings IEEE 1979 Intl. Conf. Acoustics, Speech and Signal Proc., pp. 789-792.

G. S. Kang and S. S. Everett, 1982
"Improvement of the Narrowband Linear Predictive Coder, Part I: Analysis Improvements," NRL Report 8645, Dec. 1982. AD A124313

G. S. Kang and L. J. Fransen, 1982
"Second Report of the Multirate Processor (MRP) for Digital Voice Communications," NRL Report 8614, Sept. 1982. AD A120591

J. E. Luck, 1969
"Automatic Speaker Verification Using Cepstral Measurements", Jour. Acoust Soc. Am., vol. 46, pp. 1026-1032.

C. A. McGonegal, A. E. Rosenberg and L. R. Rabiner, 1979
"The Effects of Several Transmission Systems on an Automatic Speaker Verification System," Bell Sys. Tech. Jour., vol. 58, pp. 2071-2087.

J. D. Markel, 1978
"Research on Voice Authentication," Speech Comm. Res. Lab., Inc., Santa Barbara, CA, Feb. 1978. AD A104316

J. D. Markel, B. T. Oshika and A. H. Gray, Jr., 1977
"Long-Term Feature Averaging for Speaker Recognition," IEEE Trans. Acoustics, Speech and Signal Proc., vol ASSP-25, pp. 330-337.

MIL-STD-188-113
"Common Long Haul/Tactical Standards for Analog-to-Digital Conversion Techniques."

N. Mohankrishnan, M. Shridhar and M. A. Sid-Ahmed, 1982
"A Composite Scheme for Text-Independent Speaker Recognition," Proceedings IEEE 1982 Intl. Conf. Acoustics, Speech and Signal Proc., pp. 1653-1656.

J. D. Mosko, K. N. Stevens and G. R. Griffin, 1983
"Interactive Voice Technology: Variations in the Vocal Utterances of Speakers Performing a Stress-Inducing Task," Naval Aerospace Medical Res. Lab., Pensacola, FL, Aug. 1983. AD A135932

H. Ney, 1981
"Telephone-Line Speaker Recognition Using Clipped Autocorrelation Analysis," Proceedings IEEE 1981 Intl. Conf. Acoustics, Speech and Signal Proc., pp. 188-192.

L. L. Pfeifer, 1977
"Feature Analysis for Speaker Identification," RADC-TR-77-277, Final Tech. Rept. for Rome Air Development Ctr., Aug. 1977. AD A044311

S. Pruzansky and M. V. Mathews, 1964
"Talker-Recognition Procedure Based on Analysis of Variance," Jour. Acoust. Soc. Am., vol. 36, pp. 2041-2047.

A. E. Rosenberg, 1973
"Listener Performance in Speaker Verification Tasks," IEEE Trans. Audio Electroacoustics, vol. AU-21, pp. 221-225.

A. E. Rosenberg, 1976(a)
"Automatic Speaker Verification: A Review," Proceedings IEEE, vol. 64, pp. 475-487.

A. E. Rosenberg, 1976(b)
"Evaluation of an Automatic Speaker-Verification System Over Telephone Lines," Bell Sys. Tech. Jour., vol. 55, pp. 723-744.

M. R. Sambur, 1973
"Speaker Recognition and Verification Using Linear Prediction Analysis," Quart. Prog. Rept. on ONR Contract N00014-67-A-0204-0069, Jan. 1973. AD A007502

M. R. Sambur, 1976(a)
"Speaker Recognition Using Orthogonal Linear Prediction," IEEE Trans. Acoustics, Speech and Signal Proc., vol. ASSP-24, pp. 283-289.

M. R. Sambur, 1976(b)
"Text Independent Speaker Recognition Using Orthogonal Linear Prediction," Proceedings IEEE 1976 Intl. Conf. Acoustics, Speech and Signal Proc., pp. 727-729.

A. Schmidt-Nielsen and K. R. Stern, 1985
"Identification of Known Voices as a Function of Familiarity and Narrowband Processing," Jour. Acoust. Soc. Am., vol 77, pp. 658-663, Feb 1985.

R. Schwartz, S. Roucos and M. Berouti, 1982
"The Application of Probability Density Estimation to Text-Independent Speaker Identification," Proceedings IEEE 1982 Intl. Conf. Acoustics, Speech and Signal Proc., pp. 1649-1652.

B. G. Secrest and R. E. Helms, 1977
"Remote Terminal Speaker Verification", Final Tech. Rept. for Rome Air Dev. Ctr., Report No. RADC-TR-77-169, May, 1977. AD A040827

M. Shridhar and M. Baraniecki, 1979
"Accuracy of Speaker Verification via Orthogonal Parameters for Noisy Speech," Proceedings IEEE 1979 Intl. Conf. Acoustics, Speech and Signal Proc., pp. 785-788.

G. M. Siegel and H. L. Pick, Jr., 1974
"Auditory Feedbank in the Regulation of Voice," Jour. Acoust. Soc. Am., vol. 56, pp. 1618-1624.

T. E. Tremain, 1982
"The Government Standard Linear Predictive Coding Algorithm: LPC-10," Speech Technology, vol. 1(2), pp. 40-49, April 1982.

W. D. Voiers, 1977
"Diagnostic Evaluation of Speech Intelligibility," in Speech Intelligibility and Recognition, M. E. Hawley, ed., Dowden, Hutchinson and Ross, Stroudsburg, PA.

J. J. Wolf, 1972
"Efficient Acoustic Parameters for Speaker Recognition," Jour. Acoust. Soc. Am., vol. 51, pp. 2044-2056.

E. H. Wrench, Jr., 1981
"A Realtime Implementation of a Text Independent Speaker Recognition System," Proceedings IEEE 1981 Intl. Conf. Acoustics, Speech and Signal Proc., pp. 193-196.

E. H. Wrench, Jr., 1984
"Automatic Speaker Recognition System, Appendix A," ITT Defense Comm. Div., San Diego, CA, Proposal 36031 (to Naval Res. Lab., Washington, DC), Feb. 1984.

END

FILMED

7-85

DTIC